#### Joseph Crockett

**ES-207 Environmental Data Analysis**

**February 8th 2016**

#### Homework Assignment 1 and 1.5

#### Objective Statement

Technological advances in LiDAR systems have both increased the accuracy and reduced cost in recent years, however, both system cost and labor required to process data is prohibitive at the biome-scale. Using LiDAR measurements of Sierra Nevada conifers, our objective is to determine, if any, the relationship between crown spread and tree height. A positive correlaton would inform a model that utilizes less costly remote sensing systems (i.e. Landsat/ NAIP) to estimate tree height from aerial images. From such a model, standing carbon can be estimated from conifer biomass (volumes).

#### Methods:

Linear models relating tree height to crown radii and crown radii to tree height were developed from the data described below. To do so, the data were diagnosed for obvious issues, distributions and spatial extent of the observations were plotted, tests for correlation, homoscedastity, residuals, and outliers were conducted, and all code was commited to Git for review and review.

### Data

The data are LiDAR measurements of crown height (m) and radii (m) of 176101 conifers in the Sierra Nevada Mountain Range, CA, from a dataset handled by Professor Joshua Viers, UC Merced. Exact location for observations is given as x-y coordinates (m) as per the Universal Coordinate System

### Code

# Step 1: Import data  
tree\_data <- read.csv("~/Desktop/Documents/Writings/Spring 2016/ES 207/Labs/Lab1/Trees.csv",   
 header = T) #load LiDAR tree measurements into local environment data frame   
  
dim(tree\_data) #Checking formatting of frame.

## [1] 176100 5

str(tree\_data)

## 'data.frame': 176100 obs. of 5 variables:  
## $ OBJECTID: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ x : num 707921 707921 708190 708546 707569 ...  
## $ y : num 4333002 4333002 4333002 4333002 4333002 ...  
## $ z.m. : num 2.01 2.01 12.6 42.29 15.63 ...  
## $ r.m. : num 1.03 1.03 2.11 5.47 2.93 ...

anyDuplicated(tree\_data[, 2:5]) # Check for duplicated entries

## [1] 2

# The formatting seems appropriate, but I will rename column  
# names to avoid conflicts.  
colnames(tree\_data) <- c("OBJECTID", "xval", "yval", "zval",   
 "rval")  
  
save(tree\_data, file = "~/Desktop/tree\_data.rda")

# Step 2: Run Data Diognastics  
ncol(tree\_data) #running data diagnostics

## [1] 5

nrow(tree\_data)

## [1] 176100

head(tree\_data)

## OBJECTID xval yval zval rval  
## 1 1 707921.2 4333002 2.0061 1.032689  
## 2 2 707921.2 4333002 2.0061 1.032689  
## 3 3 708189.9 4333002 12.5996 2.113910  
## 4 4 708546.1 4333002 42.2903 5.471469  
## 5 5 707569.2 4333002 15.6254 2.930143  
## 6 6 708935.7 4333002 13.4347 2.908795

tail(tree\_data)

## OBJECTID xval yval zval rval  
## 176095 176095 708708.2 4334600 5.2456 2.377610  
## 176096 176096 708753.0 4334600 5.0028 1.589963  
## 176097 176097 708753.0 4334600 5.0028 1.452385  
## 176098 176098 708762.9 4334600 44.8285 2.970038  
## 176099 176099 708817.8 4334600 8.1583 2.971006  
## 176100 176100 708817.8 4334600 8.1583 2.971006

summary(tree\_data) # shows mean, max, min, median, IQR, range, etc.

## OBJECTID xval yval zval   
## Min. : 1 Min. :707062 Min. :4333002 Min. : 2.000   
## 1st Qu.: 44026 1st Qu.:707570 1st Qu.:4333401 1st Qu.: 6.155   
## Median : 88050 Median :707997 Median :4333765 Median :10.901   
## Mean : 88050 Mean :708017 Mean :4333777 Mean :13.910   
## 3rd Qu.:132075 3rd Qu.:708471 3rd Qu.:4334154 3rd Qu.:17.776   
## Max. :176100 Max. :709000 Max. :4334600 Max. :73.927   
## rval   
## Min. : 0.1847   
## 1st Qu.: 1.5021   
## Median : 2.2416   
## Mean : 2.5956   
## 3rd Qu.: 3.3191   
## Max. :12.0044

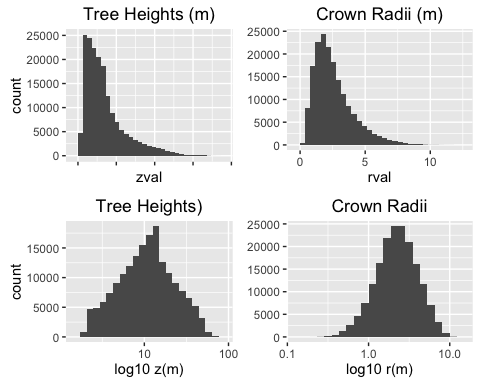
library(ggplot2) #plotting histograms using ggplot2, with a function to plot multiple graphs  
library(gtable)

## Loading required package: grid

library(grid)  
library(gridExtra)  
library(ggExtra)  
  
p1 <- ggplot(tree\_data, aes(x = zval)) + geom\_histogram() + theme(axis.text.x = element\_blank()) +   
 labs(title = "Tree Heights (m)", size = 0.5)  
  
p2 <- ggplot(tree\_data, aes(x = rval)) + geom\_histogram() + theme(axis.title.y = element\_blank()) +   
 labs(title = "Crown Radii (m)", size = 0.5)  
  
# Histograms of log transformed variables  
p3 <- ggplot(tree\_data, aes(x = zval)) + geom\_histogram(bins = 20) +   
 scale\_x\_log10() + labs(title = "Tree Heights)", x = "log10 z(m)",   
 size = 0.5)  
  
p4 <- ggplot(tree\_data, aes(x = rval)) + geom\_histogram(bins = 20) +   
 scale\_x\_log10() + theme(axis.title.y = element\_blank()) +   
 labs(title = "Crown Radii", x = "log10 r(m)", size = 0.5)  
  
grid.arrange(p1, p2, p3, p4, ncol = 2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Log transformation gives us normal distribution for tree  
# heights and crown radii.

Pearson Correlations

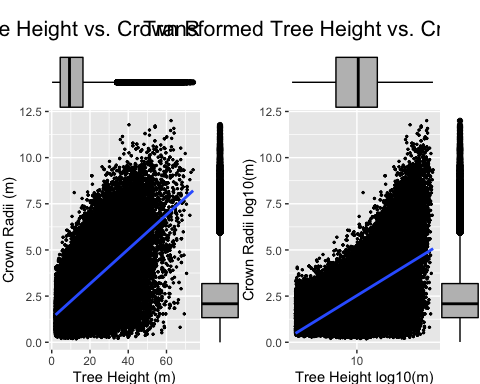
# Step 3: Examine Correlation Between Variables Is the  
# relationship linear? Comparisons of both Pearson and  
# Spearman's rho can reveal linear vs monotonic  
# relationships.  
p5 <- ggplot(data = tree\_data, aes(x = zval, y = rval)) + geom\_point(position = "jitter",   
 size = 0.5) + geom\_smooth(method = glm) + labs(title = "Tree Height vs. Crown Radii",   
 x = "Tree Height (m)", y = "Crown Radii (m)", size = 0.5)  
  
p6 <- ggMarginal(p5, type = "boxplot")

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?

p7 <- ggplot(data = tree\_data, aes(x = zval, y = rval)) + geom\_point(position = "jitter",   
 size = 0.5) + geom\_smooth(method = glm) + scale\_x\_log10() +   
 labs(title = "Transformed Tree Height vs. Crown Radii", x = "Tree Height log10(m)",   
 y = "Crown Radii log10(m)", size = 0.5)  
  
p8 <- ggMarginal(p7, type = "boxplot")

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?

grid.arrange(p6, p8, ncol = 2, nrow = 1)



# Test for correlation between Tree Height and Crown Radii An  
# assumption of Pearson's correlation is that variables must  
# be approximately normally distributed, so log  
# transformation will ensure that this assumption is held.  
  
cor.test(x = log10(tree\_data[, 4]), y = log10(tree\_data[, 5]),   
 na.action = na.omit)

##   
## Pearson's product-moment correlation  
##   
## data: log10(tree\_data[, 4]) and log10(tree\_data[, 5])  
## t = 313.23, df = 176100, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5951581 0.6011570  
## sample estimates:  
## cor   
## 0.598166

cor.test(x = tree\_data[, 4], y = tree\_data[, 5], na.action = na.omit)

##   
## Pearson's product-moment correlation  
##   
## data: tree\_data[, 4] and tree\_data[, 5]  
## t = 372.61, df = 176100, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6613456 0.6665688  
## sample estimates:  
## cor   
## 0.6639653

# Without assuming normality, we can use Spearman's rho test  
# to test correlation between non-parametric variables  
  
cor.test(x = tree\_data[, 4], y = tree\_data[, 5], na.action = na.omit,   
 method = "spearman")

## Warning in cor.test.default(x = tree\_data[, 4], y = tree\_data[, 5],  
## na.action = na.omit, : Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: tree\_data[, 4] and tree\_data[, 5]  
## S = 3.4322e+14, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.6229113

# S < P, thus we can assume linearity.

# Step 4: creating a function  
  
EDA <- function(x) {  
 # x is a list of values  
 min\_tmp <- min(x, na.rm = T)  
 max\_tmp <- max(x, na.rm = T)  
 mdn\_tmp <- median(x, na.rm = T)  
 mn\_tmp <- mean(x, na.rm = T)  
 rnge\_tmp <- range(x, na.rm = T)  
 sd\_tmp <- sd(x, na.rm = T)  
 cv\_tmp <- (sd\_tmp/mn\_tmp) \* 100  
 print(paste("Minimum =", min\_tmp, sep = " "))  
 print(paste("Maximum =", max\_tmp, sep = " "))  
 print(paste("Median =", mdn\_tmp, sep = " "))  
 print(paste("Mean =", mn\_tmp, sep = " "))  
 print(rnge\_tmp)  
 print(paste("Standard Deviation =", sd\_tmp, sep = " "))  
 print(paste("Coefficient of Variance =", cv\_tmp, sep = " "))  
   
}

Lab 1.5

# Step 5: Subsets  
tree\_data\_limit <- subset(tree\_data, xval > 708000)  
tree\_data\_limit <- subset(tree\_data\_limit, xval < 708200)  
tree\_data\_limit <- subset(tree\_data\_limit, yval > 4334000)  
tree\_data\_limit <- subset(tree\_data\_limit, yval < 4334200)  
# subsetting 708000 < x < 708200 and 4334000 < y < 4334200  
  
# By taking this subset, we are examining in greater detail  
# the center of the study area, where we earlier noted a  
# greater concentration of observations.  
  
ncol(tree\_data\_limit) #running data diagnostics

## [1] 5

nrow(tree\_data\_limit)

## [1] 1828

head(tree\_data\_limit)

## OBJECTID xval yval zval rval  
## 115905 115905 708014.4 4334000 9.6756 2.057952  
## 115913 115913 708040.5 4334000 9.4072 0.569672  
## 115922 115922 708006.1 4334000 12.5425 2.108259  
## 115931 115931 708085.2 4334000 11.4814 2.194448  
## 115932 115932 708139.5 4334000 35.2096 0.868401  
## 115933 115933 708139.5 4334000 35.2096 1.460975

tail(tree\_data\_limit)

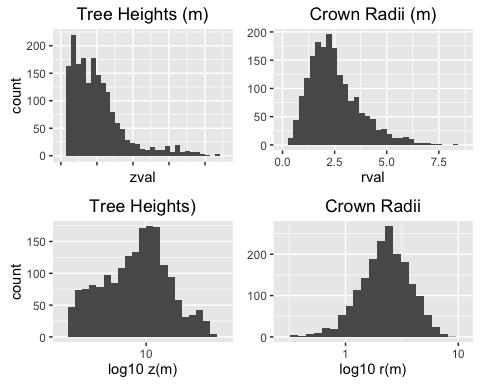
## OBJECTID xval yval zval rval  
## 135705 135705 708147.9 4334199 35.5175 6.666972  
## 135718 135718 708130.8 4334199 2.7225 1.036557  
## 135719 135719 708130.8 4334199 2.7225 1.036557  
## 135750 135750 708123.4 4334200 3.0029 0.992865  
## 135751 135751 708123.4 4334200 3.0029 0.992865  
## 135778 135778 708015.7 4334200 38.6529 7.380800

summary(tree\_data\_limit)

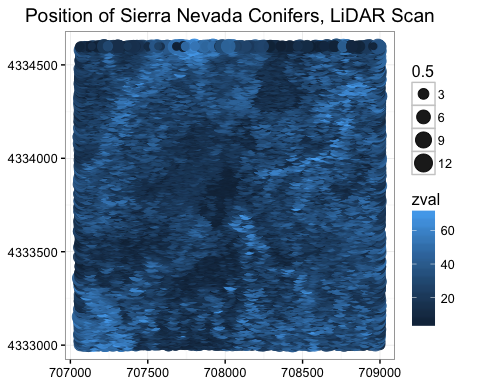
## OBJECTID xval yval zval   
## Min. :115905 Min. :708000 Min. :4334000 Min. : 2.001   
## 1st Qu.:120973 1st Qu.:708066 1st Qu.:4334061 1st Qu.: 4.702   
## Median :125254 Median :708111 Median :4334105 Median : 8.656   
## Mean :125509 Mean :708105 Mean :4334100 Mean :10.207   
## 3rd Qu.:130007 3rd Qu.:708140 3rd Qu.:4334135 3rd Qu.:13.010   
## Max. :135778 Max. :708200 Max. :4334200 Max. :43.284   
## rval   
## Min. :0.3301   
## 1st Qu.:1.6349   
## Median :2.2881   
## Mean :2.5187   
## 3rd Qu.:3.1544   
## Max. :8.2299

p9 <- ggplot(tree\_data\_limit, aes(x = zval)) + geom\_histogram() +   
 theme(axis.text.x = element\_blank()) + labs(title = "Tree Heights (m)",   
 size = 0.5)  
  
p10 <- ggplot(tree\_data\_limit, aes(x = rval)) + geom\_histogram() +   
 theme(axis.title.y = element\_blank()) + labs(title = "Crown Radii (m)",   
 size = 0.5)  
  
# Histograms of log transformed variables  
p11 <- ggplot(tree\_data\_limit, aes(x = zval)) + geom\_histogram(bins = 20) +   
 scale\_x\_log10() + labs(title = "Tree Heights)", x = "log10 z(m)",   
 size = 0.5)  
  
p12 <- ggplot(tree\_data\_limit, aes(x = rval)) + geom\_histogram(bins = 20) +   
 scale\_x\_log10() + theme(axis.title.y = element\_blank()) +   
 labs(title = "Crown Radii", x = "log10 r(m)", size = 0.5)  
  
grid.arrange(p9, p10, p11, p12, ncol = 2)

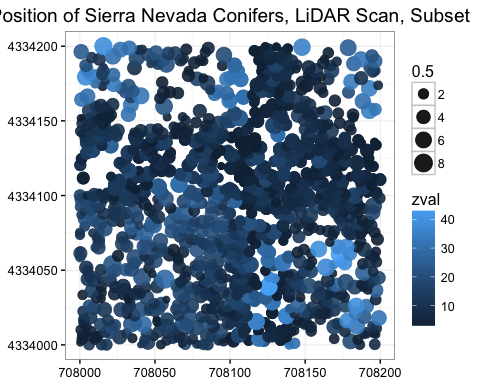
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Step 6: Spatial Plotting  
  
# Comparison of observation locations, with color indicating  
# height and dot size indicating crown radii  
ggplot(tree\_data, aes(x = xval, y = yval)) + labs(title = "Position of Sierra Nevada Conifers, LiDAR Scan",   
 x = NULL, y = NULL, size = 0.5) + theme\_bw() + geom\_point(aes(size = rval,   
 color = zval), alpha = 9/10, position = "jitter")



ggplot(tree\_data\_limit, aes(x = xval, y = yval)) + labs(title = "Position of Sierra Nevada Conifers, LiDAR Scan, Subset",   
 x = NULL, y = NULL, size = 0.5) + theme\_bw() + geom\_point(aes(size = rval,   
 color = zval), alpha = 9/10, position = "jitter")



# Step 7: Descriptive Statistics  
library(psych)

##   
## Attaching package: 'psych'

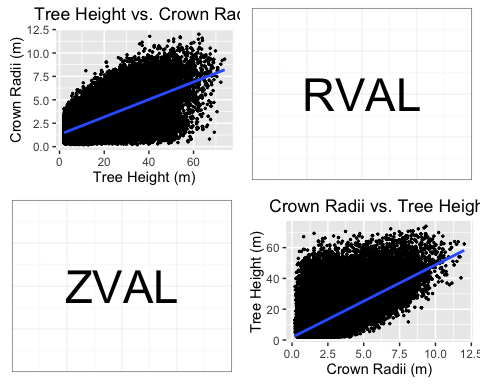
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

describe(tree\_data\_limit)

## vars n mean sd median trimmed mad  
## OBJECTID 1 1828 125509.25 5512.64 125254.50 125464.06 6714.70  
## xval 2 1828 708104.72 52.18 708111.18 708106.30 52.72  
## yval 3 1828 4334099.85 51.22 4334104.72 4334100.37 50.31  
## zval 4 1828 10.21 7.41 8.66 9.01 6.21  
## rval 5 1828 2.52 1.23 2.29 2.39 1.09  
## min max range skew kurtosis se  
## OBJECTID 115905.00 135778.00 19873.00 0.04 -1.08 128.94  
## xval 708000.05 708199.82 199.77 -0.29 -0.80 1.22  
## yval 4334000.07 4334199.99 199.92 -0.16 -0.76 1.20  
## zval 2.00 43.28 41.28 1.60 2.88 0.17  
## rval 0.33 8.23 7.90 1.02 1.15 0.03

# The subset has smaller values of mean, median, sd compared  
# to the full dataset. Standard error has increased. The  
# minimum conifer height is the same in the subset, but  
# maximum height and minimum and maximum crown radii have  
# contracted. The describe function gives the similiar  
# statistics as the summary function (min, max, etc.), and  
# includes skew, kurtosis, standard deviation, standard  
# error, trimmed mean (dropped top and bottom percentage),  
# and absolute deviation from the median. Describe does not  
# include the quartiles as summary does.

# Step 8: Regression Modeling  
  
  
p13 <- ggplot(data = tree\_data, aes(x = zval, y = rval)) + geom\_point(position = "jitter",   
 size = 0.5) + geom\_smooth(method = lm) + labs(title = "Tree Height vs. Crown Radii",   
 x = "Tree Height (m)", y = "Crown Radii (m)", size = 0.5)  
  
p14 <- ggplot(data = tree\_data, aes(x = rval, y = zval)) + geom\_point(position = "jitter",   
 size = 0.5) + geom\_smooth(method = lm) + labs(title = "Crown Radii vs. Tree Height",   
 x = "Crown Radii (m)", y = "Tree Height (m)", size = 0.5)  
  
  
p15 <- ggplot(data = tree\_data) + annotate("text", x = 708000,   
 y = 4334000, label = "RVAL", size = 12) + theme\_bw() + theme(axis.title = element\_blank(),   
 axis.ticks = element\_blank(), axis.text.y = element\_blank(),   
 axis.text.x = element\_blank())  
  
p16 <- ggplot(data = tree\_data) + annotate("text", x = 708000,   
 y = 4334000, label = "ZVAL", size = 12) + theme\_bw() + theme(axis.title = element\_blank(),   
 axis.ticks = element\_blank(), axis.text.y = element\_blank(),   
 axis.text.x = element\_blank())  
  
grid.arrange(p13, p15, p16, p14, ncol = 2, nrow = 2)



library(car)

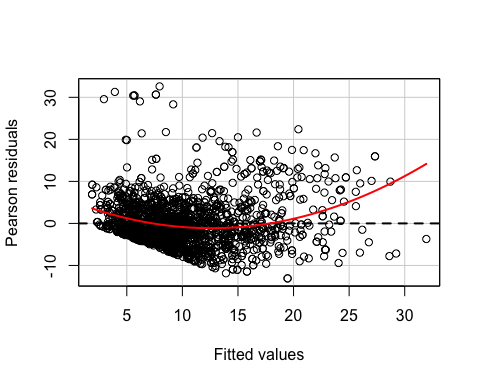
##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

tree\_car1 <- lm(zval ~ rval, tree\_data\_limit)  
tree\_car2 <- lm(rval ~ zval, tree\_data\_limit)  
  
# Residuals are the difference between the observed values  
# and the predicted values, i.e between the linear model we  
# predicted and the observation at that x value.  
  
# Height as a function of crown radii: This seems an unlikely  
# relationship, though useful if our purpose is to model  
# biomass using radii  
summary(tree\_car1)

##   
## Call:  
## lm(formula = zval ~ rval, data = tree\_data\_limit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.082 -3.659 -0.926 2.845 32.597   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6213 0.3059 2.031 0.0424 \*   
## rval 3.8058 0.1091 34.876 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.744 on 1826 degrees of freedom  
## Multiple R-squared: 0.3998, Adjusted R-squared: 0.3995   
## F-statistic: 1216 on 1 and 1826 DF, p-value: < 2.2e-16

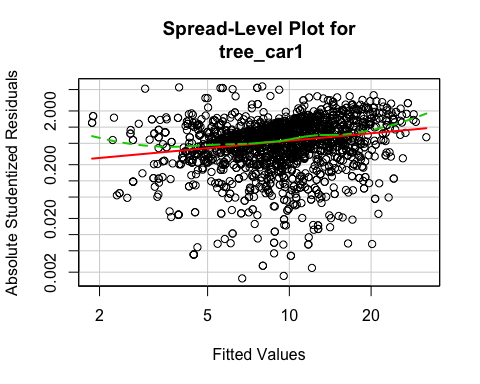
residualPlot(tree\_car1)



ncvTest(tree\_car1)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 85.69707 Df = 1 p = 2.09723e-20

spreadLevelPlot(tree\_car1)

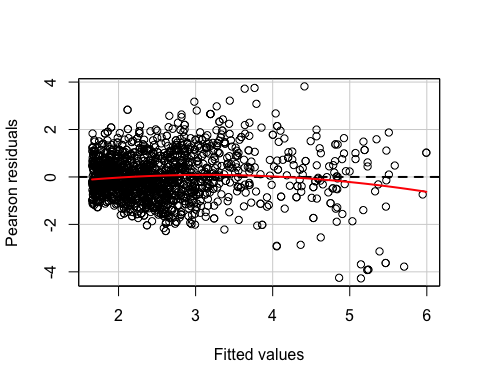


##   
## Suggested power transformation: 0.5400678

# The residual plot indicates a slightly non random  
# distribution of the residuals. Without transformation of  
# the variables, a linear model may not be applicable.  
  
# Radii as a function of tree height: a more likely causal  
# relationship.  
summary(tree\_car2)

##   
## Call:  
## lm(formula = rval ~ zval, data = tree\_data\_limit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.2769 -0.6174 -0.0414 0.5757 3.8176   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.446457 0.037993 38.07 <2e-16 \*\*\*  
## zval 0.105051 0.003012 34.88 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9543 on 1826 degrees of freedom  
## Multiple R-squared: 0.3998, Adjusted R-squared: 0.3995   
## F-statistic: 1216 on 1 and 1826 DF, p-value: < 2.2e-16

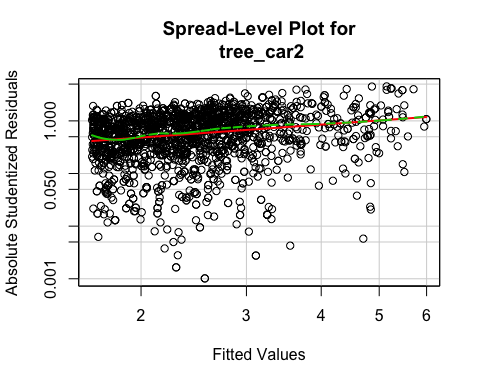
residualPlot(tree\_car2)



ncvTest(tree\_car2)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 376.7704 Df = 1 p = 6.28203e-84

spreadLevelPlot(tree\_car2)



##   
## Suggested power transformation: 0.1825716

# Homoscedasticity is the assumption that the dependant  
# variable (i.e tree height) influences with similar variance  
# the independent variable across the range of values. As we  
# can see, our subsetted data (comparable to the full  
# dataset), do not exhibit full homoscedascity, and the  
# spread level plot suggests a power transformation to  
# rectify this.

# Step 9: Outliers  
  
# Detecting outliers in both linear models  
outlierTest(tree\_car1)

## rstudent unadjusted p-value Bonferonni p  
## 120513 5.726065 1.1994e-08 2.1925e-05  
## 115932 5.493790 4.4869e-08 8.2021e-05  
## 119671 5.377938 8.5048e-08 1.5547e-04  
## 119672 5.377938 8.5048e-08 1.5547e-04  
## 119571 5.341142 1.0393e-07 1.8998e-04  
## 119572 5.341142 1.0393e-07 1.8998e-04  
## 120382 5.335954 1.0690e-07 1.9541e-04  
## 116146 5.188728 2.3530e-07 4.3014e-04  
## 115933 5.090396 3.9413e-07 7.2047e-04  
## 134187 4.965231 7.5016e-07 1.3713e-03

outlierTest(tree\_car2)

## rstudent unadjusted p-value Bonferonni p  
## 115932 -4.520714 6.5590e-06 0.011990  
## 116146 -4.490701 7.5432e-06 0.013789

# These outliers are points that may a significant influence  
# on the model despite their abnormal distance from other  
# observations. The outliers in the data could be explained  
# by equipment error; strong resource competetion in an area  
# resulting in tall trees with narrow crowns; little resource  
# competition resulting in wide crowns and relatively short  
# trunks; differences in tree species; or conifers that  
# suffered some sort of disturbance that reduced crown radii  
# or height

Step 10: Commit to Git

### Results

Plotting conifer positions reveals non-uniform distribution of trees, with high density in the center of the study area and along the southern border, close to what appears to be a path or riverbed. Distributions of tree height and crown radii are skewed right, and so we log transformed the data to better meet the assumptions of normality. Subsequent Spearman's rho and Pearson's product moment correlation tests indicated that correlation of height vs. radii exists with r-values of .623 and .598 respectively (p-value < 2.2e-16 and < 2.2e-16). Tests of residuals and homoscedasticity indicate that without transformation, neither a linear model describing crown radii as a function of tree height nor a model describing tree height as a function will adequately model the relationship (suggested power transformations are .54 and .18 respectively)

### Discussion

Our results indicate that modeling eith tree height as a function of crown radii or crown radii as a function of tree height would not completely describe the relationship. Ideally, we would test a nonlinear model or include more variables to build a multiple regression model.

### Limitations

The data used is a dense sample of conifers in the Sierras; however, it only describes a km 2 area. If our purpose is to model the relationship between Sierra Nevada conifers height and crown radii, observations from a larger spatial area would capture variations in terrain that may affect that relationship.